汉英动词次范畴化对应类型的统计分析

韩习武[1] 赵铁军[2]
(黑龙江大学计算机科学技术学院 哈尔滨 150080)[1]
(哈尔滨工业大学计算机科学技术学院 哈尔滨 150001)[2]

摘 要 基于大规模句子级,对齐双语语料库进行了统计分析汉英动词次范畴化对应类型的系统性实验。首先以语言学量度为启发,应用双重最大似然检验的统计过滤方法初步估计了654种汉英次范畴化对应类型的概率分布;然后根据汉英句法特点对次范畴化对应类型进行了语言学分类;最后针对每一种对应类型及其背景语料进行了基于支持向量机的语言学类别标注和统计可靠性分析。

关键词 汉英动词次范畴化,统计分析,支持向量机
中图法分类号 TP181 **文献标识码** A

Statistical Analysis for Chinese-English Verb Subcategorization

HAN Xi-wu[1] ZHAO Tie-jun[2]
(School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China)[1]
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)[2]

Abstract Based on large scale Chinese-English parallel corpus, this paper described a systematic experiment of statistical analysis for bilingual verb subcategorization. Firstly, with lexical and grammatical compatibility as heuristics, probabilistic distributions of 654 bilingual subcategorization frames were estimated by means of a two-fold MLE filtering method. Then, linguistic classification of the frames was determined according to Chinese and English syntax. Finally, linguistic classes for each frame were labeled via SVM on the basis of their supporting corpus.

Keywords Chinese-English verb subcategorization, Statistical analysis, SVM

关于单语动词次范畴化统计分析的研究已经在英、汉、德、捷克、西班牙、葡萄牙、希腊等语种中取得了很大程度的进展[1,2],并在不同程度上建立了有助于单语信息处理的次范畴化词汇知识库。但是在多语、跨语信息交流日益频繁的今天,世界上跨语言的次范畴化理论研究仍然很少,并且不成体系;更没有系统地自动分析双语或多语次范畴化知识的实践性研究,或提出相关研究方案。英语和汉语都属于世界上最具影响力的语种,英汉相关的信息处理量极大,并且二者在单语动词次范畴化理论研究和自动获取方面都有着相对较好的科研成果,这为开展汉英双语动词次范畴化的统计分析研究提供了良好的基础。

我们在汉英动词次范畴化统计分析的实验中采用了句法描写的框架形式。汉语直接采用此前研究[3]的138个纯句法描写的基础类型;英语采用82个次范畴化句法基础类型,其中77个是通过手工重新组合Korhonen[1]的138个句法、语义混合描写类型而获得的,另外5个在Korhonen的集合中没有对应,是根据真实语料结合语言学句法规范总结获得的。

本文在以上形式描写的基础上,基于65万汉英句对的语料库,进行了汉英跨语言次范畴化对应类型统计分析的系统性实验。文章第2节以语言学量度为启发,应用双重最大似然检验的统计过滤方法初步估计了654种汉英次范畴化对应类型的概率分布;第3节根据汉英句法特点对次范畴化对应类型进行了语言学分类;第4节针对每一种对应类型和背景语料进行了基于支持向量机的语言学类别标注和统计可靠性分析;最后总结全文并指出此后的可能研究方向。

1 对应类型的概率分布估计

实验语料包括65万句子级对齐的汉英双语句对,其中一部分来自公开免费的网络资源,另一部分是我们在翻译工作中积累下来的。为了便于谓词识别和次范畴化获取,分别应用Collins的英语中心词驱动句法分析器[4]和哈尔滨工业大学机器智能与翻译实验室的汉语中心词驱动句法分析器[5]对语料中英语和汉语句子进行了预处理。

1.1 识别潜在句对

首先用词汇兼容性和句法兼容性等语言学量度作为启发信息来识别可能包含跨语言次范畴化对应关系的汉英句对。用于计算词汇兼容性量度的双语词典共包括212,367个汉英单词,它们被组合到43,820个同义词互译词条组中。通过查双语词典的方法定义一个句对的跨语言词汇兼容性量度 C_L 如下:

到稿日期:2009-04-30 返修日期:2009-07-17 本文受国家自然科学基金(60773069,60873169)资助。
韩习武(1972-),副教授,主要研究方向为自然语言处理,E-mail:hxw@hlju.edu.cn;赵铁军(1962-),教授,CCF会员,主要研究方向为自然语言处理。

$$C_L=\frac{2\times\text{在双语词典中出现的词对数}}{\text{句对的全部单词数}}$$

用于计算句法兼容性量度的语言学单位包括名词、代词、动词、形容词、副词等跨语言普适性较强的5个句法范畴，该量度 D 如下：

$$D=\sum_{i=1}^{n}\lambda_i\frac{\text{Min}(|GE_i|,|GC_i|)+1}{\text{Max}(|GE_i|,|GC_i|)+1}$$

其中，GE_i 表示英文句法范畴，$|GE_i|$ 表示该范畴在英文句子中出现的次数；GC_i 是相应的汉语句法范畴；n 是句法范畴的类别数，它们在本次任务中起到不同程度的作用；λ_i 是相关类别的权重，由一个简单梯度下降算法在10,000句对手工分析的样本语料库上训练得到。

我们应用最大似然检验统计过滤方法接受 $C_L\geqslant 0.84$ 或 $D\geqslant 0.79$ 的句对为可能包含双语次范畴化对应关系的实验语料。对5,000句对样本的手工分析表明，本方法的精确率达到92.33%，召回率达到54.78%，最终从65万句对的语料库中识别了201,062潜在句对。

1.2 概率信息的获取和估计

典型的次范畴化自动获取系统一般包括4个组成部分，即预处理，经常涉及到统计句法分析；句法模式提取，用来分析相关句法范畴的论元类型和中心词；假设生成，把句法模式同基础类型进行匹配；统计过滤，依据统计可靠性对特定谓词的假设类型集合进行检验。

对于汉语句子，我们应用文献[3]的基于规则的分析器来提取谓语动词管辖范围内的句法论元的组合模式，该分析器识别论元类型的例(token)精确率为86.5%。对于英语句子，我们手工分析了文献[1]附录A给出的180个例句和部分典型语料，归纳了66条句法论元识别和论元组合规则，并应用该规则进行句法模式提取。对1,000句英文句子的开放语料的测试结果表明，英语句法论元类型识别结果的例精确率约为92.6%。

本文采用基于本体论元假设的启发式假设生成方法。把由内容词汇作中心词的必选论元定义为本体论元。假设是：源语言中本体论元在统计意义上比非本体论元更容易通过直接翻译而在目标语言中存留下来。表1给出了4种汉英句法论元的跨语言本体对应关系，OA_i 被定义为一种跨语言本体论元。在跨语言假设生成时，应用英语假设中的本体论元作为启发信息来指导汉语假设的生成。首先，定义跨语言本体论的元相关系数如下：

$$OAC=\frac{|\{OAs\_in\_H_E\}\cap\{OAs\_in\_H_C\}|}{|\{OAs\_in\_H_E\}\cup\{OAs\_in\_H_C\}|}$$

其中，H_E 和 H_C 分别表示英语和汉语的假设。然后，对于汉语假设生成器放松限制，同时生成多个假设，即n-best假设集。最后，依据跨语言本体论元信息，选择n-best集合中同英语假设相关系数最高的作为相应的汉语假设，即：

$$H_C=\text{Arg max}(OAC_i)=$$

$$\text{Arg max}\frac{|\{OAs\_in\_H_E\}\cap\{OAs\_in\_H_{Ci}\}|}{|\{OAs\_in\_H_E\}\cup\{OAs\_in\_H_{Ci}\}|}$$

应用这种启发式的方法，使得双语假设生成的例精确率达到了67.88%。

表1 汉英跨语言本体论元类型

OA_s	英语	汉语
OA_1	NP	NP,BAP,BIP
OA_2	AP,AS-AP,DP,PASS-VP,AS-PASS-VP	JP
OA_3	VP,TO-VP,WH-TO-VP,VPING,AS-VPING	VP
OA_4	SS,AS-IF-SS,WH-SS	SS

在统计过滤阶段，本文采用了跨语言双重最大似然过滤的方法。对文献[3]的单语方法加以修改，使其适合跨语言任务。新方法应用同样的汉语句式转换信息做启发，对汉英双语的假设进行两次过滤。算法描述如下。

输入：汉英跨语言假设$\langle escf_i,cscf_i\rangle$，估计过滤 p，及其过滤阈值 θ_1 和 θ_2；

输出：具有统计可靠性的汉英跨语言对应类型集合 S；

操作：对于每个包含英语类型 $escf_i$ 的跨语言假设，

1) 如果 $p(escf_i,cscf_i)>\theta_1$，则接受该假设到集合 S；

2) 否则，如果 $p(escf_i,cscf_i)>\theta_2$，$p(cscf_i|cscf_j)>0$，并且$\langle escf_i,cscf_j\rangle\in S$，也接受该假设到集合 S；

3) 对没有通过检验的假设重复本算法，直至集合 S 不再增大。

这里 $escf_i$ 指一个英语假设，$cscf_i$ 指一个汉语假设，$p(cscf_i|cscf_j)>0$ 表明 $cscf_i$ 和 $cscf_j$ 属于汉语转换句式。

把第2.1节识别的潜在句对用于概率信息的获取和估计，精确率为87.6%，召回率为81.35%。总共获取了654个汉英动词次范畴化对应关系的基础类型。在近25万句对的实验语料上，这些基础类型的归一化概率估计分布在0.0001到0.0746之间，句法兼容性量度大多数都超过了0.5。

2 次范畴化对应类型的语言学分类

基于次范畴化形式描写和句式转换理论，对实验语料的深入分析表明，汉英双语次范畴化对应基础类型在更大粒度上可以分成8个类别，包括4个基本类别和4个复合类型。

2.1 基本类别

(1)完全对应类型：汉英句子的句法论元分布结构几乎完全相同；这种句对包含完全对应的谓语动词，基本上是逐词对应的翻译结果；这种情况在用于次范畴化对应关系自动获取的语料库中约占19.52%[1]；

(2)转换对应类型：在一种语言方面，原本完全对应的SCF类型被其转换句式所替代，但原文的基本意义并没有因此而改变；这种句对也包含完全对应的谓语动词，在用于次范畴化对应关系自动获取的语料库中约占2.55%；

(3)派生对应类型：一种语言中的某些本体论元类型在另一种语言中由非本体论元来表示；这种句对往往包含扩展对应的谓语动词，但扩展出来的成分被重新组合到其它论元之中；这种情况在用于次范畴化对应关系自动获取的语料库中所占比率最高，约27.32%；

(4)扩展对应类型：在一种语言的SCF实现形式中增加了一些非本体论元，或由某个本体论元扩展出两个独立的本体论元；这种句对也包含扩展对应的谓语动词，并且增加的成分独立构成论元，往往用来补充说明谓语动词；这种情况在用于次范畴化对应关系自动获取的语料库中约占25.65%。

2.2 复合类别

复合类别是由转换、派生和扩展3个基本类别以任意方

[1] 这一比率是从自动标注了对应关系大类别的语料库中估计得到的，具体标注方法将在本文第3节介绍。

式组合而成的。从组合内容的角度看，可以分为派生转换、扩展转换、扩展派生以及扩展派生转换4个复合类别；从组合方式的角度看，还可以分为“与”复合与“或”复合。“与”复合的对应关系基本类别发生在同一句对当中，“或”复合的对应关系基本类别则发生在不同句对当中。

(1)派生转换复合：这种情况在用于次范畴化对应关系自动获取的语料库中约占6.39%。“与”派生转换复合是指在一种语言的句子里，原本是派生对应的SCF类型被其转换句式所替代；“或”派生转换复合是指某些原本是“与”派生转换的关系类型，在有些句对中只发生了派生对应，而转换对应的发生概率非常小或没有发生；

(2)扩展转换复合：这种情况在用于次范畴化对应关系自动获取的语料库中约占4.54%。“与”扩展转换复合是指在一种语言的句子里，原本是扩展对应的SCF类型被其转换句式所替代；“或”扩展转换复合是指某些原本是“与”扩展转换的关系类型在有些句对中只发生了扩展对应，而转换对应的发生概率非常小或没有发生；

(3)扩展派生复合：这种情况在用于次范畴化对应关系自动获取的语料库中约占11.98%。“与”扩展派生复合是指在一个句对的次范畴化对应关系基础类型中同时发生了扩展和派生两种现象；“或”扩展派生复合是指同样的对应关系基础类型在别的句对中只发生了扩展或派生的某一种；

(4)扩展派生转换复合：这种情况在用于次范畴化对应关系自动获取的语料库中所占比率最小，约2.04%。“与”三重复合是指在一个句对的次范畴化对应关系基础类型中同时发生了扩展、派生和转换3种现象；“或”三重复合是指同样的基础类型在另一些句对中只发生了一种或两种对应。

3 对应类型的类别标注和统计可靠性分析

在类别标注工作中，一方面要保持标注类别的计算语言学理论，即以句法描写为主的英汉双语动词次范畴化形式规范和相关的句式转换理论上的一致性；另一方面要保持这些类别在大规模真实语料上的统计可靠性。因此，我们采取了手工分析标注和基于机器学习方法相结合的工作方案。

3.1 手工分析标注

根据上一节的分析结果，英汉动词次范畴化对应类型分别属于8个标注类别。为每个类别从实验语料中抽取了600个较具代表性的英汉句对，共4,800句对，用于手工标注工作。每个句对的手工标注过程严格遵循如下算法：

输入：类别①完全对应、②转换对应、③派生对应、④扩展对应和一个汉英句对；

输出：标注了次范畴化对应关系类别的汉英句对；

操作：初始化当前已观察类别集合 T 为空，即 $T=\emptyset$；

1) 若句对中包含对译内容的论元成分在形式上能够按固定位置相互对应，则标注类别{①}，算法终止；否则，置 $T=\{①\}$，转下一步；

2) 若汉语句子的某一转换句式同英文句子存在 T 对应，则根据 T 进行判断：

a) 若 $T=\{①\}$ 则标注{②}，算法终止；

b) 若 $T\neq\{①\}$，则标注 $T/\{①\}\cup\{②\}$，算法终止；

否则，若 $T=\{①\}$，则转下一步，若 $T\neq\{①\}$ 则标注 T，算法终止；

3) 若英汉句子任何一方有论元成分的增减，且增减论元不影响原内容和形式均对应的论元成分的位置，则置 $T=T/\{①\}\cup\{④\}$；无条件转下一步；

4) 若有内容对应但形式不对应的论元成分，则置 $T=T/\{①\}\cup\{③\}$；转第二步。

3.2 基于SVM的自动标注

英汉句对的次范畴化类别标注问题可以看作是对输入句对按照对应谓语动词的不同次范畴化特征而进行的分类工作。应用SVM方法来完成这项分类任务，首先要构造适合英汉双语次范畴化描写方式的向量空间。根据形式化描写规范，双语次范畴化对应关系由相应谓语动词前后的句法性论元成分决定。因此，我们取相关句对的谓语动词及其前后的部分词性、词组标记和论元标记来构造这一空间，具体形式如图1所示。

$$\begin{pmatrix} C_{-3},C_{-2},C_{-1} & C_1,C_2,C_3,C_4,C_5 \\ P_{-3},P_{-2},P_{-1}\ \boxed{V_e}\ P_1,P_2,P_3,P_4,P_5 \\ B_{-3},B_{-2},B_{-1} & B_1,B_2,B_3,B_4,B_5 \end{pmatrix} \begin{pmatrix} C_{-3},C_{-2},C_{-1} & C_1,C_2,C_3,C_4,C_5 \\ P_{-3},P_{-2},P_{-1}\ \boxed{V_c}\ P_1,P_2,P_3,P_4,P_5 \\ B_{-3},B_{-2},B_{-1} & B_1,B_2,B_3,B_4,B_5 \end{pmatrix}$$

图1 汉英动词次范畴化的向量空间

该向量由英语和汉语的次范畴化框架两大部分组成。其中，$\boxed{V_e}$ 是英语谓语动词，$\boxed{V_c}$ 是汉语谓语动词，它们不记入SVM特征空间；C_{-i} 和 C_i 分别为谓语动词之前和之后的第 i 个单词的词性；P_{-i} 和 P_i 分别为谓语动词之前和之后的第 i 个词组的标记；B_{-i} 和 B_i 分别为谓语动词之前和之后的第 i 个论元成分的标记，这实际上构成了一个 3×16 的特征矩阵。为了便于计算，在SVM具体应用时把这个矩阵平铺成包含48个特征的一维向量。

此前手工标注的4,800句对可以作为支持语料，再增加1,200句对不包括任何类型次范畴化对应关系的语料，以便在最大程度上减小潜在句对识别时遗留下来的噪音。这样一来，目标类别共包括9个：8个对应关系类别，1个是非对应的噪音类别。分别为每个对应关系类别取500句对，为噪音类别取1,000句对，用作SVM的训练语料；剩下的1,000句对作为标注性能的测试语料。此次实验应用开源工具LIBSVM[2)]，应用Pairwise多分类功能来实现英汉句对的对应关系类别标注。多次实验表明，多项式核函数更适合标注任务。表2列出标注性能，其中 P 为精确率，R 为召回率，F 为综合测评指标。

表2 多项式核SVM的标注性能

标注	精确率%	召回率%	F值
①	98.99	98	98.49
②	94.17	97	95.56
③	93.2	96	94.58
④	91.51	97	94.19
②③	91.09	92	91.54
②④	93.68	89	91.28
③④	84.47	87	85.72
②③④	84.16	85	84.58
非对应	92.59	87.5	89.97
完全对应	91.37	92.63	92
全部样本	91.6	91.6	91.6

可见，完全对应类别①的标注效果最好，三重复合类别②

2) LIBSVM是一个开源的软件包，可以免费从 http://www.csie.ntu.edu.tw/~cjlin/处获得。

③④的标注效果最差。但就对应样本和全部样本上的综合指标来看，SVM类别标注的整体性能还是可以接受的。

3.3 对应关系基础类型的类别标注

基于支持向量机的分类方法在以上类别标注任务中已经取得了较好的效果，但还不能认为相应基础类型必然包含标注类别的对应方式。原因有两个：①句对的类别标注结果还不是100%正确；②句对的次范畴化对应关系基础类型是通过基于规则的方法得到的，必然由此带来更大程度的不确定性。所以，针对英汉动词次范畴化对应关系基础类型的类别标注问题，采用较高阈值 $\theta = 0.3$ 的最大似然假设检验的方法来保证标注类别的统计可靠性。具体算法如下：

输入：某一英汉动词次范畴化对应关系基础类型的全部支持句对集合 S；

输出：该基础类型的标注类别 X；

操作：

1) 应用SVM自动标注每一个句对的对应关系类别；

2) 从 S 中删除那些标注为“非对应”的句对，得到 S'；

3) 在 S' 上对每一个出现的标注类别 C_i 统计相对频率 f_{Ci} = [C_i 的出现次数]/$|S'|$；

4) 取 $X = \bigcup_{f_{Ci} \geqslant \theta} f_{Ci}$ 为当前基础类型的标注类别。

若某一次范畴化对应关系基础类型的支持句对为 m 个，且其中 n 个句对被SVM标注为“非对应”类别，根据统计学习理论，以上算法标注该类型为某一类别的统计可靠程度将不低于 $1-(1-84.16\%)^{0.3*(m-n)} = 1-(0.1584)^{0.3*(m-n)}$。当 $m-n \geqslant 10$ 时，类别标注的统计可靠程度将接近于1。

结束语 本文基于大规模语料对汉英动词次范畴化对应类型进行了统计分析。首先自动识别出那些可能包含跨语言次范畴化关系的句子对，然后通过启发式方法和双重过滤的假设检验方法初步估计了654种汉英次范畴化对应类型的概率分布，最后在语言学分类的基础上对每一种对应类型和背景语料进行了基于支持向量机的类别标注和统计可靠性分析。

此外，本文研究的基本定义、获取方法和实验规模等方面，还都有待于进一步调整、扩大和改进。并且，英文句式转换信息必然会提高双语次范畴化的分析性能；关于特定谓词对的跨语言次范畴化统计信息尚有待研究；应用动词的语义分类信息也可能发现更多的双语次范畴化对应类型。

参 考 文 献

[1] Korhonen A. Subcategorization Acquisition[D]. Trinity Hall University of Cambridge, 2001

[2] Han Xiwu, Zhao Tiejun, Qi Haoliang, et al. Subcategorization Acquisition and Evaluation for Chinese Verbs[A]//Proceedings of the COLING 2004[C]. 2004: 723-728

[3] 韩习武. 汉语动词次范畴化自动获取技术的研究[D]. 哈尔滨：哈尔滨工业大学，2005

[4] Collins M. Head-Driven Statistical Models for Natural Language Parsing[D]. University of Pennsylvania, 1999

[5] 曹海龙. 基于词汇化统计模型的汉语句法分析研究[D]. 哈尔滨：哈尔滨工业大学，2006

[6] Chang Chih-Chung, Lin Chih-Jen. LIBSVM: a library for support vector machines[EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001

(上接第226页)

参数设置为：种群规模60，交叉概率1.0，变异概率0.1。程序的终止条件为找到最优解或迭代次数达到设定的值(对于new，迭代次数设为1000，其他设为200)。

从表1可以看到，对于全空的数独new以及难度等级为easy，medium的数独，算法在100次运算中都求出了最优解。对于难度等级为hard和evil的数独，算法虽然不能做到100%求出最优解，但是100次独立运行中求出了最优解的次数超过70次。测试结果表明了算法的有效性。

表1 算法测试结果

难度等级	找到最优解的次数	最大迭代次数/时间(秒)	最小迭代次数/时间(秒)	平均迭代次数/时间(秒)
easy	100	20/7	9/3	13.6/4.7
medium	100	39/14	17/7	27.3/10.4
hard	90	43/19	28/13	33/15.3
evil	70	68/27	30/15	40.9/18.9
new	100	452/2	177/1	255.6/1.4

为了考察局部搜索在算法中的作用，程序的终止条件应设为找到最优解或迭代次数达到20000，屏蔽局部搜索，其余设置不变，得到的测试结果如表2所列。

表2 不含局部搜索的算法测试结果

难度等级	得到最优解的次数	最大迭代次数/时间(秒)	最小迭代次数/时间(秒)	平均迭代次数/时间(秒)
easy	60	19257/31	1480/3	11619/18
medium	5	5174/8	——	5174/8
hard	0	——	——	——
evil	0	——	——	——
new	80	13663/18	2190/3	6268/9.4

从表2可见，虽然迭代次数扩大了100倍，但是在100次测试中找到最优解的次数却明显地减少了，尤其是对于难度等级为hard以及evil的数独，最优解一次也没有找到。这说明局部搜索不仅使得算法可以在迭代次数较小时找到最优解，更重要的是提高了找到最优解的几率。

结束语 为了求解数独难题，首先将其转化为一个组合优化问题。然后，提出了一种在编码、初始化、交叉、变异、局部搜索等方面具有特点的遗传算法来求解它。实验结果表明，对于各种难度等级的数独问题，算法都是有效的。

参 考 文 献

[1] 孟庆铃. 数独问题人工解法的程序实现[J]. 甘肃科技，2006(9)：150-151

[2] Yato T, Seta T. Complexity and Completeness of Finfing Solution and Its Application to Puzzles[J]. IEICE Trans. Fundamentals, 2003, E86-A(5): 1052-1060

[3] Lewis R. Metaheuristics can solve sudoku puzzles[J]. Journal of Heuristics, 2007, 13: 387-401

[4] Mantere T, Koljonen J. Solving and rating sudoku puzzles with Genetic Algorithm[C]//Proceedings of the 12th Finish Artificial Intelligence Conference. Sept. 2006

[5] Goldberg DE, Alleles LR. The Traveling Salesman Problem[C]//Proceedings of an International Conference on Genetic Algorithms and Their Applications. 1985: 154-159